Folding nucleus: specific or multiple? Insights from lattice models and experiments

Eugene I Shakhnovich

In this commentary, I compare and discuss different views on the nucleation mechanisms in protein folding.

Address: Harvard University, Department of Chemistry and Chemical Biology, 12 Oxford Street, Cambridge MA 02138, USA.
E-mail: eugene@diamond.harvard.edu

Folding & Design 15 December 1998, **3**:R108–R111
http://biomednet.com/elecref/13590278003R0108

Understanding the properties of the transition state ensemble (TSE) of conformations in protein folding kinetics is an important piece in the protein folding puzzle. Not surprisingly, this issue has been addressed in a number of theoretical and experimental studies over the past few decades (for recent reviews, see [1–3]). While some attempts were made to derive the TSE from equilibrium sampling [4–6], it has been suggested that the reaction coordinate for a protein folding transition is not well enough known a priori and therefore the TSE can be derived only from kinetics [1–3,7–9].

The closely related concept of a folding nucleus, inspired by the strong thermodynamic analogy between first-order transitions and cooperative protein folding, has also been studied by many authors. Many studies focused on the search for folding nuclei defined as minimal stable elements of structure the assembly of which results in subsequent rapid assembly of the native state (kinetic condition). It was pointed out by myself and colleagues [7] that such a definition corresponds to a 'postcritical nucleus' related to the first stable structures that appear immediately after the transition state is overcome. Crucial questions include: do folding nuclei indeed exist and, if yes, what are their distinctive features? The kinetic analyses carried out in [3,7,10–12] for a number of lattice model chains of different lengths and degrees of sequence design (optimization) as well as various potential sets were able to identify such folding nuclei and characterize their properties. In particular, a specific folding nucleus (SFN) scenario was found whereby passing through the transition state with subsequent rapid assembly of the native conformation requires formation of some (small) number of specific obligatory contacts (specific nucleus). According to the SFN scenario, assembly of those obligatory contacts results in rapid folding to the native state.

It is important to note the distinction between a nucleation conformation as a conformation that is kinetically committed to descend fast and barrier-free to the native state and a folding nucleus, which is a common substructure shared by all nucleation conformations or, more generally, a common distinctive feature of all nucleation conformations. Clearly each nucleation conformation (i.e. a conformation that contains a folding nucleus) always also features other, optional native contacts, besides nucleation contacts. Specific nucleus contacts appear simultaneously in nucleation conformations with high probability, however, whereas each optional contact occurs with low probability and their number and location in a structure may vary between nucleation conformations [2,3,7]. The SFN model as formulated above was presented previously in [7]. Indeed, it is clear that the nucleation conformations studied in [7] always contained many optional contacts — Figure 4 of [7] shows that nucleation conformations typically contained up to 24 native contacts (out of a total of 40 for the 36-mer), whereas only 8 of them were identified as nucleus. Further, Figure 9 of [7] clearly shows that the folding rate depends on the energy of contacts outside the nucleus, though not as strongly as on the energy of nucleation contacts.

Another common misstatement of the SFN model is that it was studied for only one set of potentials (MJ potentials taken from [13]). In fact, the folding nucleus in the lattice 48-mer model was found in [10] for two quite different potential sets: the MJ set and the set from Kolinski, Godzik and Skolnick (KGS) [14]. Despite the fact that the MJ potentials placed letters K, D, E and R in the interior of the 48-mer for the designed sequences whereas the KGS parameters placed L, V, F, W etc there, the nucleus was shown to be at the same location in the 48-mer structure for both sets of potentials (see Figure 1 in [10]). One more general note is that because the potential for lattice simulations is usually taken in the simplest contact form, there cannot be a direct relationship between letter notation for lattice amino acids and real ones. In particular, it does not make sense to distinguish between 'charged' lattice amino acids and 'hydrophobic' ones.

As a generalization of the SFN model, a transition state classes scenario has been discussed recently by Pande, Rokshar and colleagues [3], in which nucleus conformations may be characterized not by a single set of preferred contacts but by a very small (2–3) number of alternative specific nucleus sets of contacts (folding classes). Whether a SFN or 'few folding classes' scenario is realized may be determined by the way in which sequences are designed and optimized [3,12].

In a recent publication [9], Klimov and Thirumalai (KT) have advocated an alternative model, 'multiple folding nuclei' (MFN), to describe folding kinetics in lattice models. Although KT did not give a specific definition of what they mean by MFN (e.g. how many is 'multiple' and what are the distinctive features of any of the proposed multiple nuclei?), they suggested that their scenario is dramatically different from the SFN model. We believe that the analysis of KT is inconclusive, however, as explained below. Furthermore, in cases where comparisons of simulations presented by KT with previous work are possible, their own data are consistent with the SFN model.

According to the definition of a nucleus, a two-step procedure was adopted by myself and colleagues in [7] to search for conformations that may contain folding nuclei. The first step was the analysis of conformations that appear just prior (in time) to the native conformation. The purpose of this step is to try to detect features that are common to many conformations that occur prior to reaching the native state. At this step, putative nucleus conformations (PNCs) can be identified. In the simulations presented in [7], PNCs were searched for among conformations that were no more than 50,000 Monte Carlo steps away from reaching the native state. Under the simulation conditions reported in [7], that corresponded on average to $\delta \approx 0.02$ of the total mean folding time. This stage of the analysis is, in any case, a heuristic one, serving only to give an idea of which nucleation hypothesis should then be tested.

The second and most crucial step of the analysis is to check that PNCs proposed at the first, heuristic stage satisfy the kinetic definition of a folding nucleus, that is, that they assemble fast into the native conformation. The defining feature of the nucleus conformations is that they are the minimally folded states among the ones that are on the 'native' side of the main free energy folding barrier. Therefore, the runs that start from correctly identified nucleation conformations must steadily descend (all the way 'downhill' in energy) into the native conformation without encountering any further major free energy barriers. To check these conditions, multiple folding runs starting from PNCs should be carried out [7,8]. Only the affirmative results of this part of the analysis give grounds to claims that folding nucleus conformations (specific or multiple) have been found. This was a key part of the analysis presented in [7], where we showed that simulations that start from conformations that contained proposed nucleus contacts descend rapidly to the native state without encountering any major free energy barriers. Even more conclusive evidence for that was presented in [15], where we showed that runs starting from nucleus conformations were able to fold fast at extremely low temperatures, in contrast to folding runs starting from random coil conformations. This indicated that no significant energy barriers were encountered in the runs that started from nucleus conformations. (A sample folding trajectory that starts from a correctly identified PNC is shown in Figure 6 of [7].)

In order to understand better why a meaningful search for nucleus conformations must include the second stage — verification of the kinetic accessibility requirement — consider a scenario that is the extreme opposite to nucleation: a purely random unbiased search for the native state on an uncorrelated energy landscape [16]. In this case, any search for conformations preceding (in time) the native state (part one of the analysis) will by definition identify them as PNCs because time is continuous in simulations. The observed distribution of contact frequencies in the set of PNCs in this case will be broad and will not depend on parameter δ, which tells at which part (in time) of the trajectories the PNCs were searched for. However, PNCs identified this way will not pass the kinetic criterion test: simulations starting from any of them will obviously not result in this case in fast and steady folding to the native conformation but rather in full random search.

This teaches us two very important lessons. Firstly, the kinetic accessibility test is a key part of the search for the folding nucleus in simulations. The random search example clearly shows that conformations that appear close (in time) to the transition state are not necessarily kinetically committed to fold fast (and free of major energy barriers) to the native conformation. Failure to find such kinetically committed conformations in simulations suggests the lack of conclusive identification of nucleation mechanisms (multiple or specific). Secondly, single-exponential folding kinetics do not necessarily imply a nucleation mechanism. In fact, the unbiased random search described above as the opposite to nucleation gives rise to exponential kinetics too. Nucleation is a kinetic mechanism of first-order-like (cooperative) transitions [17] between two states (thermodynamically defined as free energy minima) that differ substantially in energy (by many kT, usually extensive in the system size). In proteins, such transitions occur in designed sequences that have a large stability gap [18–21]; however, even in random sequences of 27-mers which show no cooperative folding transition, the kinetics of folding are single-exponential [22,23]. Notably, the nucleation mechanism was suggested in [24] based on exponential kinetics in an off-lattice model for which detailed thermodynamic analysis revealed no cooperative folding [25].

Dramatically, all PNCs identified by KT for a 27-mer failed the kinetic test criterion. Indeed, in all cases where KT started simulations from any of the PNCs, mean folding times were approximately the same (in some cases much longer) as in simulations that started from random coil conformations. This suggests either that the KT method for searching for PNCs is technically flawed or

that no clear-cut nucleation mechanism (specific or multiple) was operational in their 27-mer model simulations.

It is possible that the 27-mers studied by KT did not fold via a nucleation mechanism. An indication of this is that the distribution of contact frequencies in PNCs reported by KT for 27-mers did not depend on the parameter δ. As explained above, that suggests a random search mechanism rather than nucleation. Also, the distribution of contact frequencies in the PNCs for 27-mers was quite broad (Figure 10 of [9]). The fact that short chains (27-mers) may not follow a nucleation mechanism was pointed out previously in [4]. Rather, a three-stage random search (3SRS) mechanism was found more likely to explain the observed kinetics of folding of the 27-mer chains [4]. It was also pointed out in [4] that as chains become longer, the 3SRS mechanism is likely to break down and nucleation may become a dominant kinetic mechanism.

For longer chains, the folding transition is more cooperative (see e.g. Figure 1 of [9]) and their folding kinetics follow a nucleation mechanism rather than 3SRS. In earlier publications, the specific nucleus mechanism was reported for well-designed sequences of 36-mers [7] and especially for 48-mers [10,11]. For the 36-mers that KT studied, their results are also suggestive of a specific nucleus mechanism:

1. In contrast to the 27-mer case, Figure 13 of [9] clearly shows that there are four contacts that are markedly more likely to occur in the PNCs than the remaining 36 native contacts. Strikingly, the distribution of contact probabilities presented in this figure is bimodal (see Figure 3a of the following commentary), with the four most probable contacts forming a separate group. The bimodal distribution of contact frequencies is believed to be highly suggestive of a specific nucleus mechanism [3,5].

2. The probability of finding all four high-probability contacts simultaneously in PNCs is, in fact, very high and in this case, most importantly, it does depend strongly on parameter δ. When δ changed in KT simulations from 0.2 to 0.05, the probability of finding simultaneously all four contacts in a PNC increased from 0.42 to 0.64. The trend is strong and it is likely that all four of the most probable contacts may be present in nearly 100% of PNCs defined using the same $\delta = 0.02$ as was used in [7]. We also note that in a pure MFN scenario where each of 40 contacts appears in nucleation conformations with equal probability, any four specified contacts would simultaneously appear in a PNC in only 0.01% of trajectories.

3. The location of the four most probable nucleation contacts is determined by the topology of the native structure rather than by the actual energetic strength of the interactions between amino acids forming those contacts. This is another remarkable signature of a specific nucleus mechanism, as pointed out in [10,11,26].

It turns out that KT failed to document any nucleation mechanism in either case that they studied: for 27-mers they did not find a single nucleation conformation, while for 36-mers their analysis was incomplete due to the (in our opinion surprising) failure to carry out the kinetic accessibility test. Despite this, they proceeded with the claim that the MFN scenario is supported by experiment. This assertion is based on the fact that in most cases (but not all), the protein engineering analysis returns fractional φ values [27]. The point of view that fractional φ values support the MFN scenario is not correct, as was recently explained by Serrano and coworkers [26]. These authors pointed out that a SFN scenario cannot be distinguished from alternative folding scenarios like MFN "by the presence of fractional φ values but only by the behavior of these values when protein either is stabilized or destabilized" [26]. Having made a detailed analysis of experimental data on transition states for different SH3 domains and carefully considering possible scenarios, Serrano and coworkers conclude that "passing through the transition state barrier in some proteins requires formation of a defined structure with little conformational variability in some regions as postulated by Abkevich and coworkers". Clearly this conclusion is at great variance with what KT attribute to experimental results on the folding of SH3 domains.

One of the important missions of theory in protein folding is to outline possible scenarios that real proteins may follow. This helps to create frameworks that may be useful in explaining existing experiments and in planning new ones. The specific nucleus model is an extreme but clear example of a possible scenario that certainly helps us to understand the experimental results. The major implication of the SFN model for real proteins is prediction of a small number of 'kinetically important' (nucleus) positions in protein structure. Variations in the energy of nucleus contacts resulting from mutations have a stronger impact on folding rate than similar variations in the energy of non-nucleus contacts. This was indeed observed in lattice simulations [7,10,11] and in many experiments [26–28]. The implications of the predictions from the SFN model for protein engineering experiments and evolution have been highlighted in a number of publications [10,11,27,29]. As pointed out in [11,26], the most remarkable feature of the SFN model that really distinguishes it from the MFN model is the robustness of the transition state to variations in sequence. KT claimed erroneously that the SFN mechanism implies that mutations in the nucleus would fully disrupt the nucleation mechanism. In fact, the opposite was shown in lattice simulations [10] and in experiment [26] — the nucleus in the SFN model may be quite robust with respect even to multiple mutations. The reason for this is that the nucleus location in the SFN model depends

on the topology of the native structure to a greater extent than on a particular sequence. This property of the SFN was noted earlier [10,11] and was confirmed by KT for their 36-mer model (see Figure 13 of [9]). The dramatic implication of this finding is the prediction that proteins with different sequences but similar structures may have similar folding nuclei. This prediction was verified for SH3 domains [26,30] and for cold-shock proteins [31].

Careful analysis of experimental data and many simulations, including those of KT, point out that the SFN model is likely to be a useful generic framework for thinking about the folding kinetics of small proteins that fold via a two-state mechanism. KT seem to disagree with that, suggesting rather that the MFN scenario is not clearly defined by them and by no means proven in their work [9]. However, they seem to propose a model of "multiple folding nuclei involving many contacts with some occurring on average with larger probability than the other" [9]. If this is indeed their view on the transition state of protein folding, it is only semantically different from the SFN mechanism.

References

1. Fersht, A. (1997). Nucleation mechanism of protein folding. *Curr. Opin. Struct. Biol.* **7**, 10-14.
2. Shakhnovich, E.I. (1997). Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.* **7**, 29-40.
3. Pande, V.S., Grosberg, A.Y., Rokshar, D. & Tanaka, T. (1998). Pathways for protein folding: is a 'new view' needed? *Curr. Opin. Struct. Biol.* **8**, 68-79.
4. Sali, A., Shakhnovich E.I. & Karplus, M. (1994). How does a protein fold? *Nature* **369**, 248-251.
5. Onuchic, J., Socci, N., Luthey-Schulten, Z. & Wolynes, P. (1996). Protein folding funnels: the nature of the transition state ensemble. *Fold. Des.* **1**, 441-450.
6. Sheinerman, F. & Brooks, C. (1998). Calculations on folding of segment b1 of streptococcal protein g. *J. Mol. Biol.* **278**, 439-456.
7. Abkevich, V., Gutin, A. & Shakhnovich, E. (1994). Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry* **33**, 10026-10036.
8. Du, R., Pande, V.S., Grosberg, A.Y., Tanaka, T. & Shakhnovich, E. (1998). On the transition coordinate for protein folding. *J. Chem. Phys.* **108**, 334-350.
9. Klimov, D. & Thirumalai, D. (1998). Lattice models for proteins reveal multiple folding nuclei for nucleation-collapse mechanism. *J. Mol. Biol.* **282**, 471-492.
10. Shakhnovich, E., Abkevich, V. & Ptitsyn, O. (1996). Conserved residues and the mechanism of protein folding. *Nature* **379**, 96-98.
11. Mirny, L., Abkevich, V. & Shakhnovich, E. (1998). How evolution makes proteins fold quickly. *Proc. Natl Acad. Sci. USA* **95**, 4976-4981.
12. Abkevich, V., Gutin, A.M. & Shakhnovich, E. (1998). A protein engineering analysis of the transition state for protein folding: simulation in the lattice model. *Fold. Des.* **3**, 183-194.
13. Myazawa, S. & Jernigan, R. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**, 534-552.
14. Kolinski, A., Godzik, A. & Skolnick, J. (1993). The general method for the prediction of the three-dimensional structure and folding pathway of globular proteins: application to designed helical proteins. *J. Chem. Phys.* **98**, 7420-7433.
15. Abkevich, V., Gutin, A. & Shakhnovich, E. (1994). Free energy landscape for protein folding kinetics: intermediates, traps and multiple pathways in theory and lattice model simulations. *J. Chem. Phys.* **101**, 6052-6062.
16. Bryngelson, J.D. & Wolynes, P. (1989). Intermediates and barrier crossing in a random energy model (with application to protein folding). *J. Phys. Chem.* **93**, 6902.
17. Lifshits, E.M. & Pitaevskii, L.P. (1981). *Physical Kinetics.* Pergamon Press; Oxford.
18. Bryngelson, J.D. & Wolynes, P. (1990). A simple statistical field theory of heteropolymer collapse with application to protein folding. *Biopolymers* **30**, 177.
19. Goldstein, R., Luthey-Schulten, Z.A. & Wolynes, P. (1992). Optimal protein-folding codes from spin-glass theory. *Proc. Natl Acad. Sci. USA* **89**, 4918-4922.
20. Shakhnovich, E. & Gutin, A. (1993). Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl Acad. Sci. USA* **90**, 7195-7199.
21. Shakhnovich, E. (1994). Proteins with selected sequences fold to their unique native conformation. *Phys. Rev. Lett.* **72**, 3907-3910.
22. Karplus, M., Caflish, A., Sali, A. & Shakhnovich, E. (1994). Protein dynamics: from the native to the unfolded state and back again. In *Modelling of Biomolecules Structures and Mechanisms.* pp. 69-84. Kluwer Academic; Dodrecht.
23. Gutin, A., Sali A, Abkevich, V., Karplus, M. & Shakhnovich, E. (1998). Temperature dependence of folding in a simple proteinlike model: search for glass transition. *J. Chem. Phys.* **108**, 6466-6483.
24. Guo, Z. & Thirumalai, D. (1995). Nucleation mechanism for protein folding and theoretical predictions for hydrogen-exchange labelling experiments. *Biopolymers* **35**, 137-139.
25. Guo, Z. & Brooks, C. (1997). Thermodynamics of protein folding: a statistical-mechanical study of a small all β protein. *Biopolymers* **42**, 745-757
26. Martinez, J., Pissabarro, T. & Serrano, L. (1998). Obligatory steps in protein folding and the conformational diversity of the transition state. *Nat. Struct. Biol.* **5**, 721-729.
27. Itzhaki, L., Otzen, D. & Fersht, A. (1995). The structure of the transition state for folding of chymotrypsin inhibitor 2 analyzed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260-288.
28. Riddle, N.S., *et al.*, & Baker, D. (1997). Functional rapidly folding proteins from simplified aminoacid sequences. *Nat. Struct. Biol.* **4**, 805-809.
29. Fersht, A.R. (1993). Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. *Proc. Natl Acad. Sci. USA* **92**, 10869-10873.
30. Grantchanova, V., Riddle, D., Santiago, J. & Baker, D. (1998). Important role of hydrogen bonds in the structurally polarized transition state for folding of the src Sh3 domain. *Nat. Struct. Biol.* **5**, 714-720.
31. Perl, D., *et al.*, & Schmid, F.X. (1998). Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins. *Nat. Struct. Biol.* **5**, 229-235.